手話認識のための動作の多様性に応じたHMM構造生成

†

あらまし 手話は聴覚に障害を持つ人々にとって重要な意思疎通の手段である．本報告では画像からの手話単語の自動認識手法を提案する．従来，単一構造の HMM を用いた認識手法が提案されているが手話の複雑さは単語によって異なる上，発話ごとに動きが部分的に省略・変形されることがあるため適切な学習が難しい．複数発話から共通動作を抽出し，単語動作の多様性に応じた状態遷移構造をもつ HMM の生成手法を提案し，有効性を実験で示す．
キーワード 手話認識，HMM，状態遷移構造推定，動画像処理

Estimation of HMM Topology Reflecting Various Motions for Sign Language Recognition

†

Abstract Sign language is used for communicating to people with hearing difficulties. Recognition of a sign language image sequence is difficult due to the variety of hand shapes and hand motions. The previous method for sign language recognition used Hidden Markov Models(HMMs), but the models have a fixed topology that cannot represent various motions for each word. We propose a method to construct a Hidden Markov Model(HMM) that has branches and junctions to represent a structure which is different for each word. The proposed method consists of segmentation of a motion, and construction of the topology from segments. The topology is constructed from an initial topology by modifying it. With experiments, we show the effectiveness of the proposed method.
Key words sign language recognition, HMM, topology estimation, video processing

1. はじめに

画像列からの手話認識処理は特徴抽出処理と抽出された特徴の識別処理から成る．特徴の識別手法としては手の位置や速度，形状等を特徴とし，各手話単語の特徴を学習した隠れマルコフモデル (Hidden Markov Model, HMM) を用いて識別を行う手法が広く用いられてきた [1], [2]．HMM はいくつかの状態とそれらの間の遷移構造で表され，手話認識においては各々の状態が「手を上げる」,「両手の間隔を左右に広げる」等の単一の意味のある動きに対応している．つまり，部分動きの連続関係は状態遷移構造によって表されることとなる．

手話認識における重要な問題として，単語によって動作の複雑さが大きく異なること，同じ意味の手話単語動作であっても，得られる特徴量系列が発話によって大きく異なることが挙げられる．Starner らの手法 [2] では，HMM の状態数が単語に関係なく固定されている．複雑な動作も同じ状態数のモデルを用いるため，単語動作の複雑さを反映した学習をさせるのが難しい．

川東らの手法 [3] では，各モデルの状態数が単語動作から自動的に推定されるため，単純な動作に対しては状態数の少ないモデルを，複雑な動作に対しては状態数の多いモデルを用いて学習を行える．しかし，状態数の推定に用いられる閾値は話者毎に手動で調整する必要がある．また，状態遷移構造は図 1 のような単純な構造に限定されており，同じ動きが同じようにつながった動作にしか対応できない．

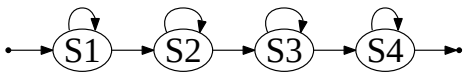


図 1 単純な線形遷移構造
Fig. 1 A linear topology

Gaolin らの手法 [4], [5] では HMM での認識処理に加えて新たな認識処理を追加することで，複数種類ある動きのうち，どれであっても許されるような状況に対応している．しかし単語に依らず状態遷移構造は固定されているため，手話単語動作内に含まれる動きの数に応じたモデルにはなっていない．

単語動作の複雑さを反映しつつ，複数種類の動きを許容する方法として，ひとつの単語に複数の HMM を対応させる方法もあるが，各単語動作の学習にはより多くのサンプルが必要となるため望ましくない．

提案法では，ひとつの手話単語に対して枝分かれを含

む状態遷移構造を持つ単一の HMM を生成する．単語動作に応じた状態遷移構造の自動生成は「動作の区間分割」と「複数発話からの区間列の統合」によって行う．まず手話単語動作を撮影した画像列を，単一の意味のある動きの区間列に分割し，各区間を状態の候補として扱う．複数の発話に渡って状態候補を収集し，それらから同じ動きと見なせるものを見つけて縮約することで状態遷移構造を生成する．

2. 学習用サンプルの区間分割

認識に用いる特徴は [3] の方法で抽出する．手の位置や形状等の特徴量は同じ話者であってもその発話毎に大きく異なるため手の動きの向きと速度を用いて区間分割を行っている．区間分割の第一段階として，手領域の重心位置の移動速度によって各フレームを静止しているフレーム，あるいは移動しているフレームに類別する．第二段階として，連続する静止フレームは静止区間としてまとめ，連続する移動フレームは，向きが揃っているものがある程度連続していれば直線動き区間としてまとめる．移動フレームではあるものの短時間の内に向きが頻繁に変わるものは振動区間として扱う．

3. 認識に用いる特徴量

一般に手話においては，手の細かい動きに重要な意味のある場合，手は顔に近い場所で動かされる傾向がある．そこで，手領域の重心位置を表す特徴量として，顔領域の重心からの相対座標 $\boldsymbol{x}$ をとり，それを以下の (1) 式で対数的に変換したものを用いる．

$$L_{r_0}\left(\left[\begin{array}{c} x \\ y \end{array}\right]\right) = \frac{\log\left(1+\frac{\sqrt{x^2+y^2}}{r_0}\right)}{\sqrt{x^2+y^2}}\left[\begin{array}{c} x \\ y \end{array}\right] \qquad (1)$$

r_0 は定数であり，ここでは初期フレームでの顔・手間の距離としている．手の位置を表す特徴量 $\boldsymbol{y}$ は，顔領域の重心位置からの相対座標 $\boldsymbol{x}$ を用いて (2) 式のようにして得る．

$$\boldsymbol{y} = L_{r_0}(\boldsymbol{x}). \qquad (2)$$

手を動かす速度は話者によって大きく異なるので，速度の値そのもので単語を識別するのは難しいが，静止しているかどうかという情報は認識に有用である．そこで手の移動速度を表す特徴量として，向きは通常の画素単位の座標系で表した速度 v に等しく，長さは平均速度の対数にした $\tilde{v}$ を用いる．

両手が接近している場合には左右の手の細かい位置関係が重要となり得るが，離れている場合には大まかな位置が重要で細かい情報はそれほど重要でない．そこで，両手の相対位置関係を表す特徴量 $\boldsymbol{y}_{\mathrm{rel}}$ は，手の重心位置を表す特徴量と同様，(1) 式を用いて対数的に変換した座標を用いる．

$$\boldsymbol{y}_{rel} = L_{r_1}(\boldsymbol{x}_{\mathrm{right}} - \boldsymbol{x}_{\mathrm{left}}) \qquad (3)$$

(3) 式の r_1 は定数であり，ここでは初期フレームでの両手間の距離としている．$\boldsymbol{x}_{\mathrm{left}}$ 及び $\boldsymbol{x}_{\mathrm{right}}$ はそれぞれ左手，右手の重心位置ベクトルである．

以上の 1 量を組み合わせた $(\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{R}}, \tilde{v}_{\mathrm{L}}, \tilde{v}_{\mathrm{R}}, \boldsymbol{y}_{rel})^T$ を特徴量ベクトルとする．

4. 状態遷移構造の生成

単一の発話から得られる区間列から初期モデルを生成し，他の発話から得られる区間列の情報を初期モデルに統合していくことで状態遷移構造の生成する．

4.1 初期モデル

同じ区間数であり区間の属性も等しいような発話は同様の動きをするグループとし，各グループ毎に初期モデルを生成する．初期モデルに含まれる状態数は区間数とし，状態のパラメータは区間に属するフレームから推定する．初期モデルの構造は図 1 のような，枝分かれのない構造となる．

4.2 HMM の統合

提案法では必要に応じて HMM の統合を繰り返し，最終的な HMM を生成する．2 つの HMM A,B を統合する処理は以下の手順から成る．

（1） HMM のそれぞれについて，対応する学習サンプルのフレームと状態との対応関係を Viterbi アルゴリズムで決定する．

（2） 割り当てられたフレームの特徴量から HMM A と HMM B の各状態の類似度を計算する．

（3） 類似度の総和が最大となるような系列を求める．

（4） この系列が状態のスキップを含んでいれば，それに対応した状態を追加する．

手順 3 において，(4) 式で表される総合的な評価値 S を最大にする対応関係 $\{S_{i(k)}, S_{j(k)}\}$ を採用する．

$$S = \sum_k C\left(S_{i(k)}, S_{j(k)}\right), \qquad (4)$$

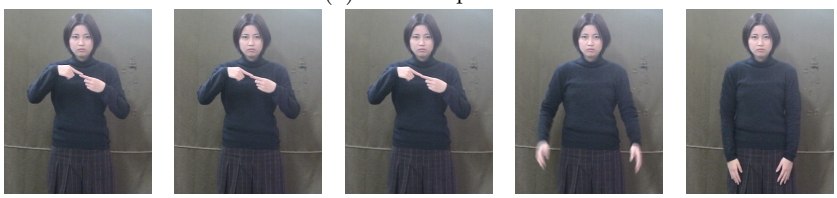
ここで S_i は HMM A の i 番目の状態，S_j は HMM B の j 番目の状態を意味している．提案法では，統合対象となる HMM の少なくとも一方は図 1 のような枝分かれのない構造を持つので，こちらについてのみ状態のスキップを許し，他方については状態遷移として辿ることのできる組み合わせ $\{S_{i(k)}, S_{j(k)}\}$ に制限し，その中で S を最大化するものを採用する．条件を満たす最良の対応関係は DP マッチングによって求められる．

状態 S_i, S_j 間の類似度としては以下のようなものを用いる．

$$C(S_i, S_j) = \hat{C}(S_i, \mathrm{samples}_j) + \hat{C}(S_j, \mathrm{samples}_i) \qquad (5)$$

ここで，$\hat{C}(S_i, \mathrm{samples}_j)$ は状態 S_j に属するサンプルと

(a) 発話 A
(a) The sample A



(b) 発話 B
(b) The sample B

time

図 2 「合う」の手話単語動作
Fig. 2 The motion of the word “match”

状態 S_i との類似度であり，次のように定義されるものとする．

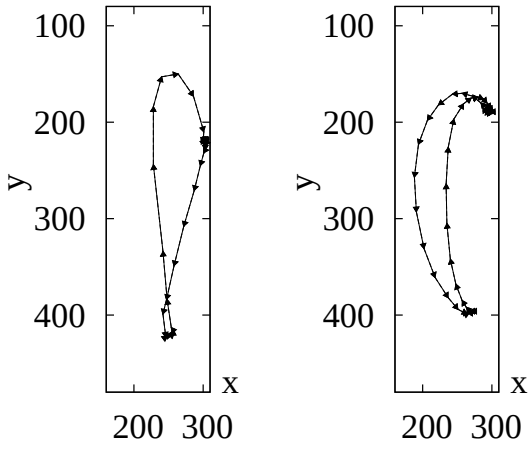
$$\hat{C}(S_i, \mathrm{samples}_j) = 1 - \frac{1}{T_j} \sum_{\boldsymbol{f} \in \mathrm{samples}_j} \left\{ \frac{d_i(\boldsymbol{f})}{\sigma} \right\}^2,$$
$$d_i(\boldsymbol{f}) = \sqrt{(\boldsymbol{f} - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\boldsymbol{f} - \boldsymbol{\mu}_i)}, \tag{6}$$

T_j は状態 S_j に属するサンプルの個数であり，$\boldsymbol{f}$ は各サンプルの特徴量ベクトルである．d_i は状態 S_i の特徴量の平均値 $\boldsymbol{\mu}_i$ と特徴量 $\boldsymbol{f}$ とのマハラノビス距離を表している．σ はマハラノビス距離の閾値を表しており，ここでは 3.0 としている．つまりマハラノビス距離が 3.0 のとき，類似度 $\hat{C}(S_i, \mathrm{samples}_j) = 0$ となる．$\hat{C}(S_i, \mathrm{samples}_j)$ が負となるような組み合わせは評価値 S を減少させるので，最良の対応関係にはそのような組は現れない．したがって σ は区間が状態と「同じ動き」と見なされるためのマハラノビス距離の閾値といえる．

続く手順 4 では，最良の対応関係 $\{S_{i(k)}, S_{j(k)}\}$ に含まれていない状態に対応した枝分かれをモデルに追加する．この統合処理によって，入力の発話の動きが含まれるよう，モデルが拡張される．

4.3 統合処理の例

提案法による状態遷移構造生成の実験例を示す．図 2 は「合う」を意味する手話単語動作を 2 名の話者が発話したときの画像列の一部である．これらの発話における右手領域の重心位置の軌跡は図 3 のようになっており，右手を左手に合わせるまでに描く弧の大きさ，合わせた後の手を戻す動作の軌跡が異なっている．このように軌跡の異なる動作のそれぞれから生成された HMM を生成し，評価値 S を最大にする対応関係を求めた結果が図 4



(a) 発話 A　　(b) 発話 B

(a)The sample A　　(b)The sample B

図 3 右手重心位置の軌跡
Fig. 3 The trajectory of right hand motions

である．手を上げる動作，両手を接触させ静止させる動作，手を下げる動作は HMM A,B の両モデルに共通している．手を上げた後と両手を接触させ静止させた後に弧を描く動作があるがその軌跡は非常に小さな弧となっている (図 3(b)) ため検出される場合と検出されない場合があり，両モデルに共通する動作とはなっていない．この対応関係から，HMM A には手を上げる動作と接触させ静止させる動作の間に現れ得る小さな弧を描く動作の可能性が反映されていないことが分かる．これを補うと図 5 のような構造が得られる．このモデルは初期モデルに採用された HMM A が表現していた動きと HMM B が表現していた動きの両方を含むモデルとなっている．

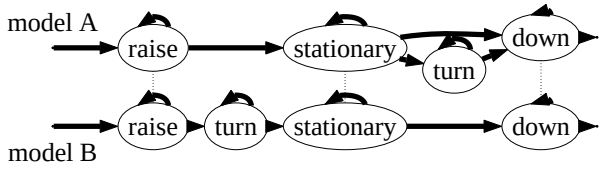


図 4　評価値 S を最大にする対応関係
Fig. 4　Matching result

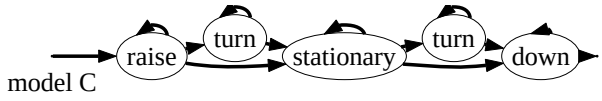


図 5　生成された状態遷移構造
Fig. 5　Generated topology

4.4　最適な HMM の選択

ある手話単語において異なるいくつかの動作が許される場合，その多様性を反映した HMM 構造には複数の候補が考えられる．例として HMM A,B で表されるような動作が許される場合について述べる．このとき両者を含む HMM 構造の候補としては以下の 4 つが考えられる．

(a)　**HMM A の構造** HMM A に対応する動作を優先して学習できる．HMM B に対応する動作の頻度が低い場合には無駄な状態を導入しないため有効である．

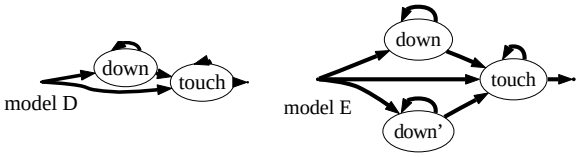
(b)　**HMM B の構造** 同上．

(c)　**HMM A と HMM B を並列に結んだ構造** 全ての学習サンプルを許容できるが，両 HMM で共通する動きがある場合にも 2 つの状態として扱われる．1 状態当たりの学習サンプル数が少なくなるため，特定のサンプルに特化したモデルになる恐れがある．

(d)　**HMM A と HMM B を統合した構造** 全ての学習サンプルを許容でき，両 HMM で共通する動きは単一の状態として扱われる．しかし 4.2 の方法では頻度の低い状態であっても削除されないため，特定の動きに特化した状態が含まれる恐れがある．

状態数が少なく構造が単純であれば 1 状態当たりの学習サンプル数を割り当てられるので，学習サンプルの総数が少ない場合にも効率的に学習できる．したがって，生成される HMM としては状態数が少なく (単純性) しかも学習サンプルに対して十分大きな尤度を与えるもの (妥当性) が望ましい．図 4 の 2 つの HMM は単純な構造であるが，この単語の全ての学習サンプルについて十分尤度が高くなるとは限らず，また図 5 の HMM は学習サンプルに対して高い尤度を与えるが単純な構造ではない．このように単純性と妥当性は trade-off の関係にある．

ここでは妥当性と単純性を総合的に評価して HMM を選択するため Minimum Description Length (MDL) 基準 [6] を用いる．これは妥当性の評価量と単純性の評価量から総合的な記述長を求め，記述長を最小とするモデルを採用する手法である．妥当性の基準は HMM のパラメータ $\boldsymbol{\theta}$ を与えたときの学習サンプル $\boldsymbol{x}$ の出力尤度



(a)before integration　　(b)after integration

図 6　HMM 候補
Fig. 6　Example of candidate HMMs

$P(\boldsymbol{x} \mid \boldsymbol{\theta})$ とし，単純性の基準には $\boldsymbol{\theta}$ を表現するのに必要な情報量を bit 単位で表した値 $L(\boldsymbol{\theta})$ を用いる．これらを用いて総合的な記述長 $L(\boldsymbol{x}, \boldsymbol{\theta})$ を以下のように定義する．

$$L(\boldsymbol{x}, \boldsymbol{\theta}) = -\log P(\boldsymbol{x} \mid \boldsymbol{\theta}) + L(\boldsymbol{\theta}), \tag{7}$$

この $L(\boldsymbol{x}, \boldsymbol{\theta})$ を最小とする HMM を採用する．

例として図 4,6 の HMM について記述長 $L(\boldsymbol{x}, \boldsymbol{\theta})$ を求めた結果を図 7 に示す．この図から，HMM C(図 5) は HMM A(図 4) に比べて複雑になっているがそれ以上に妥当性が増しており，統合によってより望ましい HMM が生成されているといえる．

図 6 の 2 つの HMM は「合う」を意味する単語動作から作成したものである．図 6(a) の HMM に線形構造を持つ HMM を統合した結果が図 6(b) の HMM である．図 7 のグラフで HMM D(図 6(a)) と HMM E(図 6(b)) を比較すると，統合によって妥当性は改善しているが，それ以上に複雑さが増しており記述量 $L(\boldsymbol{x}, \boldsymbol{\theta})$ としては統合前の HMM D の方が小さくなっている．つまり統合によって追加された状態が頻度の低い特殊な動きに対応しており，全体の性能向上にあまり寄与していないといえる．記述量 $L(\boldsymbol{x}, \boldsymbol{\theta})$ を評価することでこのような効果的でない状態を省くことができる．

4.5　モデル生成

各グループのモデルを統合して最終的な HMM を得る手順は以下の通りである．

（1）　状態数の最も少ないグループ A を選択する．

（2）　A 以外のグループで状態数最小のものを B とする．

（3）　A の HMM，B の HMM，A と B を統合した HMM の 3 つを生成し，A,B に属する学習サンプル全てを使って Baum-Welch 法による学習を行う．

（4）　3 つの HMM から MDL 基準 [7] によってひとつの HMM を選択し，A,B に属する学習サンプルと合わせて統合グループとする．

（5）　統合グループを A とし，統合されていないグループが残っていれば 2 に戻る．残っていなければ 6 へ．

（6）　統合グループの HMM をこの単語の最終的な HMM とする．

5.　認識実験結果

この実験では，2 人の話者のそれぞれに各単語につき

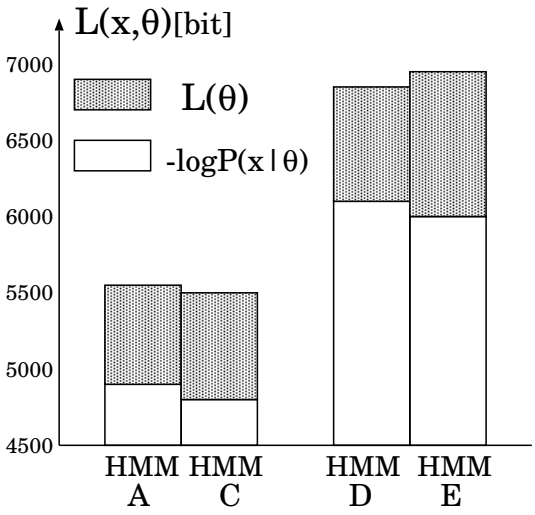


図 7 記 述 長
Fig. 7 Description length

表 1 提案法，従来法による手話単語動作の認識結果
Table 1 The number of success in recognition for words with both hands

	話者 1	話者 2
従来法 [3]	104(80.6%)	107(82.9%)
提案法	117(90.7%)	109(84.5%)

43 words, 3 motion for each word

3 回ずつ動作を行ってもらい，43 単語についての単語動作を集めた．各単語につき 1 つの発話を認識対象，残りの 5 つの発話を学習用サンプルとして認識対象のサンプルを 6 通りに変えながら認識実験を行った．比較のため，話者について手動で微調整を行う手法である川東らの方法 [3] でも同様の実験を行った．実験の結果は表 1 の通りである．提案法の認識率は 80% 以上で，手動での調整を必要とする川東らの方法に近い認識性能が得られている．提案法は手動調整によって得られた HMM に近い認識性能を持つ HMM を話者毎の調整等なしで生成できるといえる．

6. おわりに

手話単語動作を認識するための HMM を，状態遷移構造も含めて自動生成する方法を提案した．提案法は手話単語動作を意味のある単一の動きに分割する区間分割処理と，複数発話から共通する動きを見つけ縮約する統合処理から成っている．提案法を用いれば，各々の単語の持つ動きの多様性を反映したモデルを自動的に生成することができる．

文 献

[1] K. Grobel and M. Assan: “Isolated sign language recognition using hidden markov models”, Systems, Man, and Cybernetics, 1997. ’Computational Cybernetics and Simulation’., 1997 IEEE International Conference on, **1**, pp. 162–167 vol.1 (12-15 Oct 1997).

[2] T. Starner, J. Weaver and A. Pentland: “Real-time american sign language recognition using desk and wearable computer based video”, IEEE Transactions on Pattern Analysis and Machine Intelligence, **20**, 12, pp. 1371–1375 (1998).

[3] 川東, 白井, 島田, 三浦：“手話の HMM 作成のための状態分割”, 信学技報, 第 105 巻 of WIT2005-21, pp. 55–60 (2005).

[4] W. Gao, J. Ma, J. Wu and C. Wang: “Sign language recognition based on HMM/ANN/DP”, International Journal of Pattern Recognition and Artificial Intelligence, **14**, 5, pp. 587–602 (2000).

[5] G. Fang, W. Gao and D. Zhao: “Large vocabulary sign language recognition based on fuzzy decision trees”, Systems, Man and Cybernetics, Part A, IEEE Transactions on, **34**, 3, pp. 305–314 (May 2004).

[6] J. Rissanen: “A universal prior for integers and estimation by minimum description length”, The Annals of statistics, **11**, 2, pp. 416–431 (1983).

[7] J. Rissanen: “Modeling by shortest data description”, Automatica, **14**, 5, pp. 465–471 (1978).